Análisis discriminante no métrico y regresión logística en el problema de clasificación

Olga Cecilia Úsuga Manco[1]
Carmen Elena Patiño Rodríguez[2]

Resumen

Este artículo muestra los resultados de un proyecto de investigación donde se realizó un estudio de comparación entre análisis discriminante no métrico y regresión logística para el caso en el que se clasifican más de dos grupos que provienen de distribuciones normales y no normales, bajo diferentes tamaños muestrales. Este proceso se llevó a cabo por medio de un estudio de simulación, evaluando los dos procedimientos por medio de la tasa de clasificación errónea. El estudio permitió concluir que bajo distribuciones simétricas los dos procedimientos son similares en cuanto a la tasa de clasificación errónea y bajo distribuciones no simétricas la regresión logística se comporta mejor que el análisis discriminante no métrico.

Palabras clave

Análisis discriminante, análisis discriminante no métrico, regresión logística, clasificación.

[1] Docente asistente. Departamento de Ingeniería Industrial. Universidad de Antioquia. Medellín - Colombia. Calle 67 Nº. 53-108, A.A 1226. Teléfono (574) 210 55 75. Fax (574) 263 82 82. Medellín - Colombia. ousuga@udea.edu.co.

[2] Docente asistente. Departamento de Ingeniería Industrial. Universidad de Antioquia. Medellín - Colombia. Calle 67 Nº. 53-108, A.A 1226. Teléfono (574) 210 55 75. Fax (574) 263 82 82. Medellín - Colombia. cpatino@udea.edu.co

Fecha de recepción: 7 de noviembre de 2008
Fecha de aceptación: 12 de diciembre de 2008

Abstract

A study of comparison between non-metric discriminant analysis and logistic regression when the classification system has more than two clusters is shown, taking into account normal and non-normal distributions under different sample sizes. This process was carried out by means of a simulation study, evaluating the two procedures by the rate of misclassification. This study is derived from a research project and concluded that symmetrical distributions under the two procedures are similar in terms of the rate of misclassification and non-symmetrical distributions under the logistic regression performs better than the non metric discriminant analysis.

Key words

Discriminant analysis, no metric discriminant analysis, logistic regression, classification.

1. Introducción

Se han realizado estudios de comparación entre análisis discriminante y regresión logística, considerando el análisis discriminante desde el punto de vista lineal, el cual se basa en la combinación lineal de variables que separan de la mejor manera dos clases de objetos o eventos, como son los estudios realizados por Efron, Bradley (1975), Harrell, F.E., y Lee, K.L. (1985), Castrillón, F. (1998), Fan, X. y Wang, L. (1999), Pohar, *et al.* (2004), Lei, P. y Koehly, L. (2003), Richard's, *et al.* (2008). El objetivo de este artículo es presentar una comparación entre la regresión logística y un nuevo procedimiento de clasificación, como el análisis discriminante no métrico, para el caso de más de dos grupos. Según Raveh (1989), una de las ventajas del procedimiento no métrico es que es aplicable a variables cualitativas y cuantitativas y por lo tanto no tiene supuestos sobre distribuciones específicas. Hasta el momento se han realizado pocos estudios en este sentido, sólo para el caso de dos grupos como el estudio realizado por Úsuga (2006).

Usualmente estos estudios de comparación se han realizado a la luz de distribuciones como la normal, mostrando que la regresión logística obtiene buenos resultados para el caso en que las distribuciones normales son similares en cuanto a la matriz de varianzas y covarianzas y los vectores de medias. Este artículo, como resultado de un proyecto de investigación, presenta las ventajas de los dos métodos cuando se clasifican más de dos grupos que provienen de distribuciones normales y no normales, de tipo continuo.

El desarrollo de este artículo será el siguiente: la sección de materiales y métodos incluye la definición de los dos métodos de clasificación y la descripción de la generación de los datos de entrenamiento y el procedimiento de simulación; la siguiente sección incluye los resultados del procedimiento de comparación entre los dos métodos, una aplicación sencilla del procedimiento y la sección final contiene las conclusiones a la investigación.

2. Materiales y métodos

2.1 Análisis discriminante no métrico, NDA

Raveh (1983,1989) discutió y enfatizó el caso de separación de dos grupos, el cual aplicó a un conjunto de datos con múltiples grupos (3 y 10), sin definir el índice de separación para múltiples grupos. Raveh (1983) declaró que se podría realizar una generalización del coeficiente de discriminación para múltiples grupos en trabajos posteriores, pero éste fue explícitamente realizado por Guttman (1988), quien generalizó el índice de separación para múltiples grupos de Raveh y lo llamo *disco* (discriminant coefficient, coeficiente discriminante). Disco es más resistente a outliers, observaciones extremas, y es mucho más interpretable estadísticamente que el criterio de Fisher, en el caso de análisis discriminante lineal.

Sea X una matriz de datos de dimensión (nxp), en la cual los n individuos están divididos en G grupos diferentes con n_g individuos en el grupo g, es decir $n = \sum_{g=1}^{G} n_g$ y cada individuo está descrito por p variables cuantitativas. La matriz de datos puede ser considerada como una concatenación de submatrices $X(g)$ para $g = 1,...,G$, donde $X(g) = \{x_{ij}(g)\}$ es una matriz de datos de dimensión $(n_g xp)$ que describe los n_g individuos en el grupo g y $x_{ij}(g)$ designa el valor del carácter j en el individuo i que pertenece al grupo g. Sea $X_i(g) = \{x_{il}(g),...,x_{ip}(g)\}$ un vector columna que representa el individuo i en el grupo g, $\bar{x}_j(g) = \frac{\sum_{i=1}^{n_g} x_j(g)}{n_g}$ es la media del carácter j en el grupo g, $\overline{X}(g) = \{\bar{x}_{il}(g),...,\bar{x}_{ip}(g)\}$ es el vector de la media muestral total.

A continuación se presenta el coeficiente discriminante de Guttman-Raveh llamado *disco* y su interpretación.

2.1.1 Disco

Sea Z una variable aleatoria y $Z_i(g)$, un valor medido en la observación i del grupo g y sea $\bar{z}(g) = \sum_{i=1}^{n_g} \frac{z_i(g)}{n_g}$. El coeficiente discriminante de Guttman-Raveh entre G grupos está definido como:

$$disco = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g n_h \left|\bar{z}(g) - \bar{z}(h)\right|}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|z_i(g) - z_j(h)\right|} \qquad (1)$$

El numerador es $\sum_{g=1}^{G}\sum_{h=1}^{G}\left|\sum_{i=}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g) - z_j(h)\right]\right|$, donde $\sum_{i=}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g) - z_j(h)\right]$ es una medida de la separación entre los grupos h y g. El término $\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|z_i(g) - z_j(h)\right|$ del denominador representa la variación total entre los grupos h y g, medida por desviaciones absolutas. Por la desigualdad $\sum_{g=1}^{G}\sum_{h=1}^{G}\left|\sum_{i=}^{n_g}\sum_{j=1}^{n_h}\left[z_i(g) - z_j(h)\right]\right| \le \sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|z_i(g) - z_j(h)\right|$, la función *disco* satisface la siguiente desigualdad: $0 \le disco \le 1$, con $disco = 0$ indica que todas las muestras tienen la misma media y si $disco = 1$, entonces no existe traslape entre los scores de cualquiera de los dos grupos.

Sea η un vector p-dimensional, y considere la variable aleatoria $z_i(g) = \eta X_i(g)$, el cual representa el score de la observación i en el grupo g en el vector η. La ecuación (1) puede representarse como una función del vector η, de la siguiente manera:

$$disco(\eta) = \frac{\sum_{g=1}^{G}\sum_{h=1}^{G} n_g n_h \left|\eta'\left[\bar{X}_g - \bar{X}(h)\right]\right|}{\sum_{g=1}^{G}\sum_{h=1}^{G}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left|\eta'\left[X_i(g) - X_j(h)\right]\right|} \qquad (2)$$

El análisis discriminante no métrico propuesto por Raveh (1989) consiste en buscar el vector η que maximiza *disco* en (2). Éste se llama no métrico, porque, para cualquier grup h y g, tal que $\eta[\bar{X}(g) - \bar{X}(h)] > 0$, maximiza el conjunto de desigualdades $\eta[X_i(g) - X_j(h)] \ge 0$ para $i = 1,\ldots,n_g$ y $j = 1,\ldots,n_h$.

2.1.2 Regla de clasificación

Raveh (1989) describe la regla de clasificación para dos grupos. Esta regla puede ser fácilmente generalizada a G grupos de la siguiente forma. Se considera n score de *disco* $z_i(g)$ para $i = 1,...,n_g$ y $g = 1,...,G$. Se asume que los n_1 scores del grupo 1 se espera que sean menores que los n_2 scores del grupo 2, el cual se espera que sean más pequeños que los n_3 scores del grupo 3 y así sucesivamente. La regla sugerida es el punto de corte para los dos primeros grupos, 1 y 2, la cual es igual al percentil muestral 100(n_1/n) de los n scores, el punto de corte del segundo y tercer grupo es igual al percentil muestral 100($n_1 + n_2/n$) de los n scores, y así sucesivamente. Si los G grupos no tienen sobrelapamiento en sus scores, la regla de clasificación está dada por el punto de corte de los grupos g y $(g + 1)$ como:

$$\frac{(\max_{i=1,...,n_g} Z_i(g) + \min_{i=1,...n_{g+1}} Z_i(g+1))}{2},$$

el cual asegura que la tasa de error aparente de mala clasificación de las observaciones muestrales sea cero.

2.2 Regresión logística

La técnica de regresión logística multinomial, MLR, consiste en la estimación de la probabilidad de que una observación x pertenezca a cada uno de los grupos, dados los valores de las p variables que conforman la observación.

El modelo compara $G - 1$ categorías contra una categoría de referencia dadas n observaciones (y_i, x_i), donde x_i es un vector con p variables y y_i es una variable aleatoria independiente multinomial con valores $1,2,...,G$, la cual indica el grupo al cual pertenece cada observación. La probabilidad condicional de pertenencia de x_i a cada grupo está dada por:

$$P(y = j|x_i) = \frac{e^{\alpha_{1j} + \beta'_{1j} x_i}}{1 + \sum_{k=2}^{G} e^{\alpha_{1j} + \beta'_{1k} x_i}}$$

Donde $\alpha_{11} = 0$ y $\beta_{11} = 0$.

La regla de clasificación consiste en que a cada nueva observación con *p* variables se le calcula la probabilidad de que esta observación pertenezca a cada uno de los *G* grupos y luego se asigna al grupo que presentó la mayor probabilidad.

La ventaja de MLR es que no requiere supuestos distribucionales y por lo tanto se puede aplicar a distribuciones multivariadas con variables cuantitativas o cualitativas.

2.3 Datos de entrenamiento

El estudio de simulación fue llevado a cabo para comparar el poder de separación y localización del análisis discriminante no métrico y la regresión logística multinomial para el caso de más de dos grupos, en este caso de 3,5 y 7 grupos, ya que en estudios anteriores no se ha considerado la influencia de este factor en el comportamiento de los dos procedimientos. La comparación tuvo dos objetivos: a) comparar la bondad de separación por medio de la tasa de clasificación errónea para los datos dados a través de datos de entrenamiento, b) comparar la bondad de clasificación de nuevas observaciones tomadas aleatoriamente de distribuciones conocidas donde las reglas de clasificación son estimadas por medio de datos de entrenamiento. La bondad de clasificación fue medida por medio de la tasa de clasificación errónea para nuevos datos.

El análisis discriminante no métrico fue llevado a cabo por medio de una herramienta computacional como los algoritmos genéticos y la regresión logística mediante una función existente en el software estadístico R (2007).

El estudio de simulación fue llevado a cabo utilizando cuatro distribuciones: multinormal con matrices de varianza y covarianza iguales y diferentes, Lognormal, $Sinh^{-1}$ normal y Logit normal, para estas distribuciones se tomaron dos muestras de entrenamiento de tamaños 50, 100 y 200, cada una de 3, 5 y 7 variables para el caso de clasificación de 3,5 y 7 grupos.

Los datos de entrenamiento fueron obtenidos aleatoriamente por medio del software estadístico R a través de funciones como la mvrnorm de la librería MASS, *Main Package of Venables and Ripley's MASS*, y otras creadas a partir de algoritmos que generan distribuciones de probabilidad multivariada como las anteriormente indicadas.

Para la generación de una distribución multinormal con varianza de 1 y correlación entre variables de 0.5, con vector de medias μ = (0,0,0), con tamaño muestral de 100 y para el caso de 3 variables, se utilizó la función de R: mvrnorm, la cual tiene como argumentos el tamaño de la muestra, el vector de medias y la matriz de varianzas-covarianzas.

Para evaluar el desempeño de los dos procedimientos se calculó la tasa de error o probabilidad de clasificación errónea, la cual está dada por *TCE = NCE/NOBS,* donde *NCE* corresponde al número de clasificaciones erradas por la técnica de validación y *NOBS* corresponde al número de observaciones en el conjunto de validación.

El procedimiento de comparación fue llevado a cabo en cinco pasos:

1. Los datos de entrenamiento fueron seleccionados aleatoriamente de tres, cinco y siete distribuciones, Grupo 1, 2 y 3. Los grupos difirieron en sus parámetros de localización y de dispersión y tamaño de la muestra.
2. El algoritmo de NDA fue aplicado en tres datos de entrenamiento multivariados (p=3, p=5 o p=7) para encontrar la tasa de clasificación errónea.
3. Para el mismo conjunto de datos del paso 2 se aplicó la función de MLR y se encontró el número de clasificaciones erróneas.
4. Nuevos datos fueron generados aleatoriamente de la misma distribución que fue usada en el paso 1. Cada observación multivariada fue clasificada de acuerdo con el punto de corte hallado en el paso 2, para el análisis discriminante no métrico, y en el paso 3 para la regresión logística multinomial,

respectivamente. Finalmente se encontró la tasa de clasificación errónea.

5. Los cuatro pasos anteriores se repitieron 1.000 veces y se encontró la tasa promedio de clasificación errónea para el NDA y la MRL.

3. Resultados

En la tabla 1 se muestran los resultados para el caso en el que las poblaciones provienen de distribuciones normales con matrices de varianza covarianza iguales para cada uno de los grupos, con varianza de 1 y correlación entre variables de 0.5 y con tamaños muestrales de 100. Los valores de $\mu_1 = 0$, $\mu_2 = 0.5$, $\mu_3 = 1$ indican vectores, $\mu_1 = (0,0,0)$, $\mu_2 = (0.5,0.5,0.5)$, $\mu_3 = (1,1,1)$, y los valores de $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 2$ indican vectores, $\mu_1 = (0,0,0)$, $\mu_2 = (1,1,1)$, $\mu_3 = (2,2,2)$, para el caso de 3 variables.

Tabla 1. Tasa promedio de clasificación errónea para distribuciones normales multivariadas con tamaños muestrales de 100

		Parámetros de localización			
		$\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$		$\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$	
Número de grupos	Número de variables	NDA	MLR	NDA	MLR
3	3	0,5116	0,5150	0,3703	0,3672
	5	0,5033	0,5115	0,3553	0,3560
	7	0,4974	0,5107	0,3499	0,3537
5	3	0,6256	0,6128	0,4420	0,4304
	5	0,6244	0,6164	0,4000	0,3960
	7	0,6308	0,6168	0,4016	0,3844
7	3	0,6740	0,6597	0,4805	0,4680
	5	0,6971	0,6497	0,4722	0,4571
	7	0,6548	0,6391	0,4605	0,4040

Estos resultados permiten ver que en general para los parámetros planteados la regresión logística muestra mejor desempeño en el sentido de que presenta menores tasas de clasificación errónea de las observaciones. Aunque bajo el supuesto de distribuciones normales con matrices de varianza covarianza iguales las tasas de clasificación errónea son similares.

En la tabla 2 se muestran los resultados para el caso en el que se quiere clasificar 5 grupos que provienen de distribuciones normales con matrices de varianza covarianza iguales, con varianza de 1 y correlación entre variables de 0.5 y con diferentes tamaños muestrales. Los valores de $\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$ indican vectores, $\mu_1 = (0,0,0)$, $\mu_2 = (0.5,0.5,0.5)$, $\mu_3 = (1,1,1)$, y los valores de $\mu_1 = 0$, $\mu_2 = 1, \mu_3 = 2$ indican vectores, $\mu_1 = (0,0,0)$, $\mu_2 = (1,1,1)$, $\mu_3 = (2,2,2)$ para el caso de 3 variables.

Tabla 2. Tasa promedio de clasificación errónea para clasificar 5 grupos con distribuciones normales multivariadas

		Parámetros de localización			
		$\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$		$\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$	
Tamaño muestral	Número de variables	NDA	MLR	NDA	MLR
50	3	0,6000	0,608	0,4392	0,4400
	5	0,6048	0,6096	0,4256	0,4008
	7	0,6344	0,6216	0,3960	0,3744
100	3	0,6256	0,6128	0,4420	0,4304
	5	0,6244	0,6164	0,4000	0,396
	7	0,6308	0,6168	0,4016	0,3844
200	3	0,6114	0,6094	0,4488	0,4392
	5	0,6414	0,5970	0,4108	0,4022
	7	0,6312	0,6174	0,4456	0,4064

Los resultados de la tabla 2 muestran un mejor comportamiento de la regresión logística multinomial para los tres tamaños muestrales presentados.

En la tabla 3 se muestran los resultados para el caso en el que se desea clasificar 3 grupos de tamaño 100 cada uno que provienen de distribuciones normales con matrices de varianza covarianza diferentes para cada uno de los grupos, la matriz de varianzas y covarianzas para el primer grupo está definida por una varianza de 1 y correlación entre variables de 0.5 y las demás matrices son combinaciones lineales de la primera matriz. Los valores de $\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$ indican vectores, $\mu_1 = (0,0,0), \mu_2 = (0.5,0.5,0.5), \mu_3 = (1,1,1)$, y los valores de $\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$ indican vectores, $\mu_1 = (0,0,0), \mu_2 = (1,1,1), \mu_3 = (2,2,2)$ para el caso de 3 variables.

Tabla 3. Tasa promedio de clasificación errónea para clasificar 3 grupos de tamaño 100 con distribuciones normales multivariadas con matrices de varianza covarianza diferentes

Matrices de varianza covarianza	Número de variables	Parámetros de localización			
		$\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$		$\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$	
		NDA	MLR	NDA	MLR
$\Sigma_2 = 2\Sigma_1, \Sigma_3 = 3\Sigma$	3	0.5763	0.5547	0.4675	0.4418
	5	0.5566	0.5683	0.4383	0.4466
	7	0.5766	0.5433	0.4600	0.4250
$\Sigma_2 = 2\Sigma_1, \Sigma_3 = 4\Sigma_1$	3	0.5875	0.5603	0.4787	0.4446
	5	0.6183	0.5266	0.4866	0.4150
	7	0.5600	0.5283	0.4516	0.4316
$\Sigma_2 = 2\Sigma_1, \Sigma_3 = 5\Sigma_1$	3	0.5970	0.5632	0.5128	0.4509
	5	0.5733	0.5883	0.4716	0.4533
	7	0.5750	0.5500	0.4533	0.4466

Para este caso en el que se consideran observaciones que provienen de distribuciones normales con matrices de varianza covarianza diferentes la regresión logística multinomial presenta mejor desempeño que el análisis discriminante no métrico.

En la tabla 4 se muestran los resultados para el caso en el que se desea clasificar 3 grupos con 2 variables que provienen de distribuciones Lognormal aplicando un sistema de transformación sugerido por Johnson (1987) a las componentes individuales de una distribución multivariada. Los tres grupos tuvieron la misma matriz de varianzas y covarianzas, se consideraron $\sigma_1 = \sigma_2 = 1$ con diferentes valores de correlación $\rho = 0.1,0.5,0.9$.

Tabla 4. Tasa promedio de clasificación errónea para distribuciones lognormal

	Parámetros de localización					
	$\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$		$\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$		$\mu_1 = 0, \mu_2 = 2, \mu_3 = 4$	
Tamaño muestral	NDA	MLR	NDA	MLR	NDA	MLR
50	0,4680	0,4733	0,3293	0,3620	0,1446	0,1453
	0,5180	0,5166	0,3773	0,4026	0,1906	0,1886
	0,4916	0,5150	0,3566	0,3716	0,1566	0,1433
100	0,5083	0,5300	0,3750	0,3966	0,1766	0,1616
	0,5283	0,5450	0,4050	0,4150	0,2150	0,2083
	0,4825	0,5008	0,3458	0,3658	0,1466	0,1241
200	0,5050	0,5200	0,3866	0,3741	0,1433	0,1325
	0,5225	0,5416	0,4125	0,4025	0,2066	0,2275
	0,5325	0,5608	0,4091	0,4108	0,2200	0,2008

Estos resultados permiten ver que para los vectores de medias $\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$, el análisis discriminante no métrico obtuvo un mejor desempeño, aunque en los demás vectores de medias el comportamiento de los dos procedimientos fue similar.

En la tabla 5 se muestran los resultados para el caso en el que se desea clasificar 3 grupos con 2 variables que provienen de distribuciones $Sinh^{-1} - normal$, aplicando un sistema de transformación sugerido por Johnson (1987) a las componentes individuales de una distribución multivariada. Los tres grupos tuvieron la misma matriz de varianzas y covarianzas, se consideraron $\sigma_1 = \sigma_2 = 1$ con diferentes valores de correlación $\rho = 0.1,0.5,0.9$.

Tabla 5. Tasa promedio de clasificación errónea para distribuciones *Sinh⁻¹ – normal*

	Parámetros de localización					
	$\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$		$\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$		$\mu_1 = 0, \mu_2 = 2, \mu_3 = 4$	
Tamaño muestral	NDA	MLR	NDA	MLR	NDA	MLR
50	0,6026	0,5586	0,4793	0,3926	0,2306	0,1986
	0,6180	0,5333	0,5140	0,4133	0,2640	0,1840
	0,6153	0,5653	0,4860	0,4653	0,2593	0,2313
100	0,6353	0,5113	0,5266	0,3980	0,2360	0,1833
	0,6000	0,5373	0,5393	0,4420	0,3173	0,2113
	0,6440	0,5553	0,5633	0,4700	0,3040	0,2153
200	0,6290	0,5113	0,5460	0,3600	0,2870	0,1490
	0,6406	0,5333	0,5746	0,3736	0,2870	0,1736
	0,6333	0,5470	0,5903	0,4106	0,2933	0,2350

Los resultados de esta simulación permiten ver que la regresión logística multinomial tiene mejor desempeño que el análisis discriminante no métrico, además se observa que la tasa de clasificación errónea tiende a disminuir a medida que existe mayor distancia entre los vectores de medias de los grupos ya definidos.

En la tabla 6 se muestran los resultados para el caso en el que se desea clasificar 3 grupos con 2 variables que provienen de distribuciones Logit normal aplicando un sistema de transformación sugerido por Johnson (1987) a las componentes individuales de una distribución multivariada. Los tres grupos tuvieron la misma matriz de varianzas y covarianzas, se consideraron $\sigma_1 = \sigma_2 = 1$ con diferentes valores de correlación ρ = 0.1,0.5,0.9.

Tabla 6. Tasa promedio de clasificación errónea para distribuciones logit normal

	Parámetros de localización					
	$\mu_1 = 0, \mu_2 = 0.5, \mu_3 = 1$		$\mu_1 = 0, \mu_2 = 1, \mu_3 = 2$		$\mu_1 = 0, \mu_2 = 2, \mu_3 = 4$	
Tamaño muestral	NDA	MLR	NDA	MLR	NDA	MLR
50	0,5980	0,6666	0,5833	0,6846	0,5860	0,6586
	0,6006	0,6546	0,5780	0,6733	0,5753	0,6326
	0,6100	0,6553	0,5760	0,6613	0,5606	0,6346
100	0,6096	0,6770	0,5950	0,6673	0,5793	0,6773
	0,6233	0,6816	0,5800	0,6233	0,5700	0,6333
	0,6266	0,6366	0,5883	0,6183	0,6000	0,7116
200	0,6083	0,6708	0,5950	0,6441	0,5783	0,6925
	0,6116	0,6625	0,5950	0,6283	0,5950	0,6975
	0,6166	0,6391	0,6066	0,6566	0,5850	0,6950

Estos resultados permiten ver que el análisis discriminante no métrico tiene mejor desempeño que la regresión logística multinomial para el caso en que las observaciones provienen de una distribución Logit normal con los parámetros ya mencionados.

4. Aplicación

La aplicación consiste en los factores de riesgo asociados con bajo peso de bebés recién nacidos. Esta base de datos contiene información de 59 bebés que contiene las siguientes variables:

- Edad de la madre
- Peso de la madre en el último periodo menstrual
- Estado de fumar durante el embarazo
- Historia de hipertensión
- Presencia de irritabilidad uterina
- Número de controles durante el embarazo
- Peso del bebé: menos de 2.500 g, menos de 1.500 g y menos de 1.000 g

A partir de estas variables se clasificó los bebés de acuerdo con el peso en el nacimiento, como: peso bajo, peso muy bajo y peso extremadamente bajo. Cada uno de estos grupos tiene 8,13 y 38 bebés. Se encontró que disco = 0.9415. La tabla 7 muestra el número de clasificaciones correctas e incorrectas de los bebés por análisis discriminante no métrico y regresión logística.

Tabla 7. Número de clasificaciones erróneas de los bebés por NDA y MLR, en paréntesis

Grupo actual	Grupo en el cual se clasificó		
	1	2	3
1	5 (7)	2 (2)	4 (5)
2	3 (1)	9 (10)	10 (5)
3		2 (1)	24 (28)

5. Conclusiones

1. En general, la regresión logística multinomial presenta un mejor desempeño que el análisis discriminante no métrico comparando numéricamente las tasas de clasificación errónea.
2. Los vectores de medias son un factor importante en el valor de las tasas de clasificación errónea e igualmente la correlación entre las variables de las muestras generadas, ya que entre más correlación exista entre las variables la tasa de clasificación errónea puede incrementarse y si los valores de los vectores de medias son muy cercanos pueden conducir a obtener tasas de clasificación grandes.
3. Los tamaños muestrales son un factor importante en la comparación aunque los comportamientos de los dos procedimientos son similares.
4. En algunos casos las tasas promedio de clasificación errónea son bastante altas para los dos procedimientos lo cual genera inconformidad para la utilización de alguna de ellas. Se podría

plantear otra regla de clasificación diferente a la que se propone en el procedimiento de análisis discriminante no métrico.

5. Para distribuciones diferentes a la distribución normal, especialmente aquellas que no tienen relación con la distribución normal permite mostrar la ventaja de la regresión logística sobre el análisis discriminante no métrico.
6. En aquellos casos donde la tasa de clasificación errónea para los dos procedimientos es similar el análisis discriminante no métrico tiene ventaja con respecto a la regresión logística en el sentido de la interpretación de los resultados, ya que el procedimiento no métrico se interpreta en términos lineales y la regresión logística no.

Referencias bibliográficas

Castrillón, Francisco (1998). "Comparación de la discriminación normal lineal y Cuadrática con la regresión logística para clasificar vectores en dos poblaciones". Medellín. Tesis Magíster en Estadística. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín.

Efron, Bradley (1975). "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis". En *Journal of the American Statistical Association*, Nro. 70, pp. 892-898.

Fan, X. y Wang, L. (1999), "Comparing Linear Discriminant Function with Logistic Regression for the Two-Group Classification Problem", En *Journal of Experimental Education*, Vol. 67, Nro. 3, 265-286.

Guttman, L, (1998), "Eta, disco, odisco and F", En *Psychometrika*, Nro, 53, Pp, 393-405.

Harrell, F.E., y Lee, K.L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. En P.K. Sen (Ed.): B*iostatistics in biomedical: Public Health and Enviromental Sciences*. pp. 333-343. North-Holland: Elsevier Science Publishers.

Johnson, M. (1989). Multivariate statistical simulation. United States: Jhon Wiley & Sons. 230 p.

Lei, P. y Koehly, L. (2003). "Linear Discriminant Analysis Versus Logistic Regression: A Comparison of Classification Errors in the Two-Group

Case". En *The Journal of Experimental Education*, Vol. 72, Nro 1, pp. 25-49.

Pohar, M., Blas, M., y Turk, S. (2004), "Comparison of Logisitic Regression and Linear Discriminant Analysis: A Simulation Study". En *Metodoloski zvezki*, Vol. 1, Nro. 1, pp, 143-161.

Raveh, A, (1983), "Preference structure analysis: A nonmetric approach" ,en *Patter Recognition,* Nro. 16, Pp. 253-259.

Raveh, A, (1989), "A Nonmetric Approach to Linear Discriminant Analysis", En *Journal of the American Statistical Association*, Nro. 84, pp, 176-183.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Richard's, et.al. (2008), "Técnicas estadísticas de clasificación: un estudio comparativo y aplicado". En *Psicothema.* Vol. 20, Nro. 4. pp. 863-871.

Úsuga, O, (2006). 'Comparación entre análisis discriminante no métrico y regresión logística'. Tesis de grado. Universidad Nacional de Colombia, Sede Medellín.